1/1 WPIL - (C) Thomson Derwent- image
AN - 1997-303342 [28]
XP - N1997-250923
TI - Index generation system for searching document database - associates contents of first and second memory using identifier and word extracted from partial document as key
DC - T01
PA - (XERF) FUJI XEROX CO LTD
NP - 1
NC - 1
PN - JP09114856 A 19970502 DW1997-28 G06F-017/30 6p *
AP: 1995JP-0290408 19951012
PR - 1995JP-0290408 19951012
IC - G06F-017/30 G06F-017/21 G06F-017/27
AB - JP09114856 A
The system has a divider (3) which divides the input document into some divisions. An identifier provision part provides the identifier providing unit corresponding to each divided document part. A first memory (4) matches and stores the identifier and position information of each partial document part.
- A word is extracted from each partial document part by an extraction part (6). A second memory (7) matches and stores the extracted word with the corresponding identifier of each partial document part. An index generation unit (8) generates an index used as a keyword, by associating the information stored in the first and second memory.
- ADVANTAGE - Enables reduction of index data size by avoiding overlapping and describing position information of same word, thus reducing memory capacity. Improves document search processing speed. (Dwg.1/5)
MC - EPI: T01-J05B3 T01-J11D
UP - 1997-28

Search statement 32

?ST Y

Session finished: 22 JUL 2002 Time 07:12:44

WPIL - Time in minutes : 6.57
The cost estimation below is based on Questel's standard price list

Estimated cost : 17.85 USD
Records displayed and billed : 33
Estimated cost : 49.47 USD
Cost estimated for the last database search : 2[illegible].02 USD
Estimated total session cost : 1[illegible] USD

QUESTEL - Time in minutes : 0,07
The cost estimation below is based on Questel's standard price list

Estimated cost : 0.06 USD

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平9－114856

(43)公開日　平成９年(1997)５月２日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/413	３１０Ｂ
	17/27			15/20	５５０Ｆ
	17/21				５７０Ｎ
					５９０Ｅ
				15/40	３７０Ａ

審査請求　未請求　請求項の数２　ＦＤ　（全６頁）

(21)出願番号　特願平7－290408

(22)出願日　平成７年(1995)10月12日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号

(72)発明者　山浦　富久美
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内

(72)発明者　舘野　昌一
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内

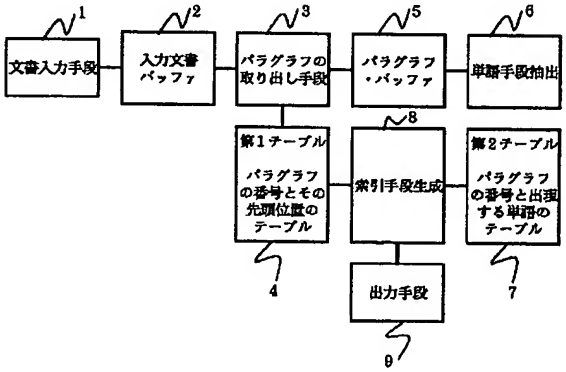
(74)代理人　弁理士　守山　辰雄

(54)【発明の名称】　検索用索引生成装置

(57)【要約】

【課題】　文字数（バイト数）で表現される部分文書の位置情報を同一の語について重複して記述することを回避して、従来に比してデータ量を大幅に減少させた索引を生成する。

【解決手段】　パラグラフ取出し手段３によって入力された文書を複数の部分文書に分割するとともにこれら部分文書に一意の識別子を付与し、第１テーブル４に部分文書の識別子と当該部分文書の文書中における位置情報とを対応付けて格納する。一方、単語抽出手段６によって部分文書から抽出された語を、当該語を抽出した部分文書の識別子と対応付けて第２テーブル７に格納する。そして、索引生成手段８が第１テーブル４と第２テーブル７との格納情報を部分文書の識別子によって関連付けて、語をキーとした索引を生成する。



1

【特許請求の範囲】

【請求項１】　文書中における検索対象の語の位置を、当該文書を構成する部分文書の位置により記述した索引を生成する装置において、

文書を複数の部分文書に分割する分割手段と、

分割された部分文書に一意の識別子を付与する識別子付与手段と、

部分文書の識別子と当該部分文書の文書中における位置情報とを対応付けて格納する第１記憶手段と、

部分文書から語を抽出する抽出手段と、

語を抽出した部分文書の識別子と抽出された語とを対応付けて格納する第２記憶手段と、

第１記憶手段と第２記憶手段との格納情報を部分文書の識別子によって関連付けて語をキーとした索引を生成する索引生成手段と、

を備えたことを特徴とする検索用索引生成装置。

【請求項２】　請求項１に記載の検索用索引生成装置において、

分割手段は予め設定された基準に従って文書を分割することを特徴とする検索用索引生成装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、電子文書データベースを検索するに際して用いられる索引を生成する装置に関し、特に、全文データベースの索引を生成するに適した装置に関する。

【０００２】

【従来の技術】学術文献、書籍、新聞等の電子文書に関するデータベースには、著者や標題等の書誌的事項だけを収録した索引型データベース、書誌的事項に抄録を加えた抄録型データベース、本文自体を収録した全文データベース等、種々の形式があるが、近年では、情報量の豊富な全文データベースが広く利用させている（「情報処理」１９９２年４月（Ｖｏｌ．３３，Ｎｏ．４）第４１３頁～第４２０頁、「情報処理」１９９２年１０月（Ｖｏｌ．３３，Ｎｏ．１０）第１１４４頁～第１１５３頁）。

【０００３】データベースに収録されている情報を検索するに際して、当該情報に含まれる語をキーとした索引を利用することにより、検索処理を迅速に行うことができる。特に、全文データベースにおいては、実用的な処理時間で検索を実現するために、索引を用いて検索することが必須とも言える。一般的に、索引には、文書中に出現する語（文字列）と、文書を構成している部分文書の当該語を含むものの文書中における位置が記述されており、この部分文書の位置として、１つの文書を対象とする場合には文書の先頭文字から当該部分文書の先頭文字までの文字数（バイト数）、複数の文書を対象とする場合にはこれら文書識別子と文書の先頭文字から当該部分文書の先頭文字までの文字数（バイト数）との組が用

2

いられている。なお、部分文書とは、文書を句点、段落等で区切った文書の構成部分をいう。

【０００４】図５には、従来の索引の一例を示してある。この索引は、同図（Ａ）に示すテーブルと同図（Ｂ）に示すテーブルとから成っている。（Ａ）に示すテーブルには、或る文書から抽出した語５０（出現する語）と、当該語５０を（Ｂ）のテーブルに関連付けるポインタ５１とが含まれており、（Ｂ）に示すテーブルには前記語５０を含む部分文書の文書中における位置５２が含まれている。語の位置５２には各語毎の文書中における出現位置をまとめて記述してあり、これら位置５２は文書の先頭文字からこれら語の先頭文字までの文字数（バイト数）でそれぞれ表現されている。例えば、”春”という語は、文書中の５６バイト目から始まる部分文書、文書中の１２０バイト目から始まる部分文書、・・・・に位置していることが記述されている。

【０００５】

【発明が解決しようとする課題】上記したように、従来の索引にあっては、文書中から抽出された多数の語に対してそれぞれ部分文書の位置を記述し、これら部分文書の位置を文書の先頭からの文字数（バイト数）で表現していた。すなわち、同一の部分文書中に含まれる互いに異なる語についてもそれぞれ同一の部分文書の位置情報が格納され、これらの位置情報が文字数（バイト数）で表現されていた。このため、索引全体ががかなり大きなデータ量のものとなってしまい、索引を格納するためのメモリの大型化によるコスト増大を招くばかりか、索引に基づく検索処理の遅延化も招くという問題があった。

【０００６】本発明は上記従来の事情に鑑みなされたもので、文字数（バイト数）で表現される部分文書の位置情報を同一の語について重複して記述することを回避して、従来に比してデータ量を大幅に減少させた索引を生成する装置を提供することを目的とする。また、これによって索引を小型化し、コスト低減及び検索処理の迅速化を実現する索引生成装置を提供することを目的とする。

【０００７】

【課題を解決するための手段】上記目的を達成するため、本発明の検索用索引生成装置では、文字数（バイト数）で表現される位置情報に比べて小さなデータ量となる識別子で部分文書を記述し、文書から抽出した各語にはそれぞれ部分文書の識別子を対応付ける。そして、これら部分文書の識別子には文字数（バイト数）で表現される部分文書の位置情報を対応付け、これによって、部分文書の位置情報を同一の語に対して重複して記述することを回避する。そして、各識別子は各部分文書に一意に対応していることから、各識別子に部分文書の位置情報が重複して対応付けられることもなく、結果として、索引のデータ量が従来に比して大幅に減少する。

【０００８】すなわち、分割手段によって文書を複数の

部分文書に分割して、識別子付与手段によって分割された部分文書に一意の識別子を付与し、第１記憶手段に部分文書の識別子と当該部分文書の文書中における位置情報とを対応付けて格納する。一方、抽出手段によって部分文書から抽出された語を、当該語を抽出した部分文書の識別子と対応付けて第２記憶手段に格納する。そして、索引生成手段が第１記憶手段と第２記憶手段との格納情報を部分文書の識別子によって関連付けて、語をキーとした索引を生成する。

【０００９】なお、文書を部分文書に分割する位置は、構造化文書におけるタグのように文書中に明示的に記しておいてもよいが、そういった記号がなくても、文書のパラグラフ（節、段落等）を単位とする、文字列の或る長さ（Ｎバイト）以内のもっとも長い文章の切れ目（句点）を単位とする、文字列の或る長さ（Ｎバイト）以上の最も短い文章の切れ目を単位とする、といったように種々設定することができる。

【００１０】

【発明の実施の形態】本発明に係る検索用索引生成装置を実施する場合の一形態を図面を参照して説明する。図１に示すように、検索用索引生成装置は、索引化する文書データを入力するための文書入力手段１と、入力された文書データを一時記憶する入力文書バッファ２と、一時記憶された文書データをパラグラフ毎の部分文書データに分割するとともに各部分文書データに一意のパラグラフ番号（識別子）を付与するパラグラフ取出し手段３と、付与されたパラグラフ番号と当該部分文書データの位置情報とを対応付けて格納する第１テーブル４と、分割された部分文書データを一時記憶するパラグラフバッファ５と、一時記憶された部分文書データからキーとなる語を抽出する単語抽出手段６と、抽出された語と当該語を抽出した部分文書データのパラグラフ番号とを対応付けて格納する第２テーブル７と、第１テーブル４と第２テーブル７との格納情報から抽出された語をキーとした索引を生成する索引生成手段８と、生成された索引を二次記憶装置等に出力して格納する索引出力手段９と、を備えている。

【００１１】文字列入力手段１は全文データベースに収録する文書を検索用索引生成装置に読み込むための手段であり、通常テキストデータとして与えられる文書データを入力文書バッファ２内に格納する。本実施例では部分文書の単位を１パラグラフとしており、パラグラフ取出し手段３は入力文書バッファ２内に記憶された文書データをパラグラフ毎の部分文書データに分割し、各部分文書データを順次パラグラフバッファ５に格納する。

【００１２】また、この分割処理に際して、パラグラフ取出し手段３は各部分文書データに一意の識別子を付与するものであり、本実施例では部分文書が文書中の先頭から数えて何パラグラフ目かを示すパラグラフ番号を各部分文書データに付与する。更に、この分割処理に際して、パラグラフ取出し手段３は各部分文書の先頭の文字が文書の先頭の文字から数えて何文字目かをカウントし、カウントされた文字数（バイト数）で表現される各部分文書の位置情報を検出する。

【００１３】第１テーブル４は読み出し書き込み自在なメモリから構成されており、パラグラフ取出し手段３によって得られた各部分文書毎のパラグラフ番号と位置情報（バイト数）とを対応付けて記憶する。例えば、図３の（Ａ）に示すように、パラグラフ番号”１”の部分文書はバイト数”０”の位置から始まり、パラグラフ番号”２”の部分文書はバイト数”５６”の位置から始まり、パラグラフ番号”１０２”の部分文書はバイト数”８６０２０”の位置から始まるといったように、各部分文書毎のパラグラフ番号と位置情報とを対応付けて記憶する。

【００１４】単語抽出手段６は、パラグラフバッファ５に一時記憶された部分文書データを形態素解析してキーとなる語（例えば、自立語）を抽出し、抽出した語と当該部分文書データのパラグラフ番号とを対応付けて第２テーブル７に格納する。なお、部分文書データから語を抽出するためには、形態素解析以外（例えば、ＤＰマッチング法等）の手法を用いることもできる。第２テーブル７は読み出し書き込み自在なメモリから構成されており、抽出された語とパラグラフ番号とを対応付けて記憶する。例えば、図３の（Ｂ）に示すように、パラグラフ番号”１”の部分文書から”古来”、”日本人”、”四季”、”変化”、”生活”、”一部”、”ある”という語が抽出された場合には、これらの語をパラグラフ番号”１”で示されるメモリ領域にまとめて格納する。なお、第２テーブル７は第１テーブル４と別途のメモリ装置から構成してもよいが、同一のメモリ装置に領域を分割して第１テーブル４とともに構成してもよい。

【００１５】索引生成手段８は、上記のように第１テーブル４と第２テーブル７とに格納されら情報をパラグラフ番号によって互いに関連付け、これによって、単語抽出手段６によって抽出された語をキーとした索引を生成する。すなわち、索引生成手段８は、第２テーブル７の格納情報に基づいて、図４の（Ａ）に示すように抽出された語４０毎にまとめたテーブル１１を作成するとともに、同図の（Ｂ）に示すように各語４０に対応しているパラグラフ番号４２を各語毎にまとめたテーブル１２を作成し、テーブル１１の各エントリ４０（各語）をそれぞれポインタ４１でテーブル１２の各エントリ４２（パラグラフ番号群）に対応付け、更に、テーブル１２のパラグラフ番号群４２を第１テーブル４の格納情報（同図の（Ｃ））に対応付ける。

【００１６】上記構成の検索用索引生成装置によると、従来に比してデータ量が大幅に減少した索引が以下のようにして生成される。まず、図２に示すテキスト文書が文書入力手段１から入力されると、この文書データがバ

ッファ２に格納される。そして、バッファ２に格納された文書データに対して、パラグラフ取出し手段３が分割処理を行って、部分文書データを順次取り出し、更に、各部分文書データにパラグラフ番号を付与するとともに各部分文書データの位置情報（バイト数）をカウントする。

【００１７】これら各部分文書データのパラグラフ番号と位置情報（バイト数）は第１テーブル４に格納される一方、各部分文書データはパラグラフバッファ５に一時記憶されて、当該部分文書データから単語抽出手段６によってキーとなる語が抽出される。そして、これら抽出された語は当該語を抽出した部分文書のパラグラフ番号とともに第２テーブル７に格納され、第１テーブル４と第２テーブル７との格納情報に基づいて索引生成手段８によって図４に示す索引が生成される。

【００１８】すなわち、パラグラフ取出し手段３によって、入力文書の１番目のパラグラフ「古来より日本人にとって四季の変化は生活の一部であった。」が取り出されて、バッファ５に格納されるとともに、このパラグラフ番号”１”の部分文書の先頭位置は文書中の０バイト目なので、図３の（Ａ）に示すように、第１テーブル４のパラグラフ番号”１”のエントリには０が書き込まれる。そして、単語抽出手段６によって、バッファ５に格納された部分文書から「古来」、「日本人」、「四季」、「変化」、「生活」、「一部」、「ある」の自立語が抽出され、これらの語が第２テーブル７にパラグラフ番号”１”の部分文書に出現する語として登録される。

【００１９】上記と同様にして、入力文書中の２番目のパラグラフ「春はあけぼの。夏は・・・・・」が取り出され、パラグラフ番号”２”の部分文書の先頭位置は文書の先頭から５６バイト目なので、第１テーブル４の２番目のエントリには５６が書き込まれる。このように、第１テーブル４のＮ番目のエントリの値はＮ番目のパラグラフの文書の先頭からのバイト数を表している。そして、２番目のパラグラフから「春」、「あけぼの」、「夏」、「夜」・・・・といった語が抽出され、パラグラフ番号”２”の部分文書における出現語として第２テーブル７に登録される。入力文書中の３番目以降の各部分文書についても同様な処理が繰り返しなされ、第１テーブル４及び第２テーブル７に所定の情報か登録される。

【００２０】このようにして第１テーブル４と第２テーブル７とが作成された後、索引生成手段８が第２テーブル７の語を重複を排して文字コード順に並べ換え、図４の（Ａ）に示す形式のテーブル１１を生成する。また、テーブル１１の各語を抽出した部分文書のパラグラフ番号を各語毎にまとめて図４の（Ｂ）に示すテーブル１２を生成し、テーブル１１の各語をポインタ４１で関連付けるとともに、テーブル１２の各パラグラフ番号を第１テーブル４のパラグラフ番号に対応付ける。すなわち、索引生成手段８によって図４の（Ａ）、（Ｂ）、（Ｃ）に示す各テーブル１１、１２、４がパラグラフ番号で関連付けられ、語をキーとした１つの索引が生成される。

【００２１】このように生成された索引は、図５に示した従来の索引に比べて、データ量（バイト数）の多くなる部分文書の位置情報５３が重複していない形式となっており、従来に比してデータ量の少ないコンパクトなものとなっている。更に、同一パラグラフ内にある同じ語（例えば、パラグラフ番号”２”の部分文書中にある「春」）は、パラグラフ番号を１つテーブル１２に記憶すればよいので、テーブル１２のエントリ数を削減することができ、更に索引のデータ量が少なくて済むようになっている。

【００２２】上記のようにして生成された索引を用いて検索を行うときには、与えられた検索語によってテーブル１１を検索して該当する語を探し、該当する語のエントリのポインタ４１が示すテーブル１２のエントリを参照して、その検索語が現れる部分文書（パラグラフ）の先頭文字の文書中における位置（なお、次のエントリの値との差から部分文書の長さも）を得ることができる。そして、検索結果として、該当する部分文書の先頭部分等をディスプレイ装置等に表示する。

【００２３】なお、検索語はの分書中における出現位置は部分文書の位置として得られ、検索語の詳しい位置がわからないので、語と語の位置関係（例えば、”特許”と”出願”が隣り合って現れる、”特許”の後２０文字以内に”出願”が現れる等）といった検索は索引を用いた処理だけでは行うことができない。しかしながら、索引から検索語の有無及びその検索語が含まれる部分文書が特定できるので、その部分文書のテキストそのものを参照して検索語と検索語との位置関係を確認すれば、所期の目的を容易に達成することができる。また、部分文書を１パラグラフや１ページといったような小さな単位に設定すれば、そのテキストを参照して検索語同士の位置関係を確認する作業は、あまり時間をとらずに容易に行うことができる。

【００２４】なお、部分文書の単位の設定に、パラグラフ、セクション、節、章といった階層をもたせ、生成された索引の利用範囲を拡大してもよく、このように部分文書単位の階層を表現する場合にあっても、索引全体から見れば容量の小さなテーブル１２を拡張するだけであるので、従来に比して索引をコンパクトなものに維持することができる。

【００２５】

【実施例】約３２０Ｋｂｙｔｅの入力文書データについて、上記した本発明に係る装置で索引を生成したところ、パラグラフ数は約１０００、文書中に出現する異なる語の数は約６７００、図４の（Ｂ）に示すテーブル１２のエントリ数は約２７０００、１エントリは２ｂｙｔ

ｅであり、同図の（Ｃ）に示すテーブル４のエントリ数は約１０００、１エントリは３ｂｙｔｅであった。すなわち、これらテーブル１１、４の容量はそれぞれ、５４Ｋｂｙｔｅ、３Ｋｂｙｔｅ程度であった。

【００２６】一方、図５に示した従来技術による索引では、上記のテーブル１１及び４に該当するテーブル（同図の右側部分）のエントリ数は約５９０００、１エントリは３ｂｙｔｅとなり、当該テーブルの容量は約１７７Ｋであった。したがって、従来と同一なテーブル１１（図４の（Ａ））を除いて比べると、従来では１７７Ｋｂｙｔｅ必要であったものに対し、本発明によれば約１／３の５７Ｋｂｙｔｅで済み、索引全体としても大幅にデータ量が削減されたことが確認された。

【００２７】

【発明の効果】以上説明したように、本発明に係る検索用索引生成装置によると、文書中に含まれる語の位置を特定するための部分文書を一意に付与した識別子によって識別して、バイト数表現される部分文書の位置情報を同一の語に対して重複して記録することを回避するようにしたため、従来に比してデータ量を大幅に減少させて索引を生成することができ、索引の小型化によって必要＊＊とするメモリ容量を減少させてコスト低減を達成することができ、更には、索引を用いた検索処理の迅速化を達成することができる。

【図面の簡単な説明】

【図１】　本発明の一例に係る検索用索引生成装置を示す構成図である。

【図２】　索引を生成するための入力文書の一例を示す図である。

【図３】　第１テーブル及び第２テーブルに格納される情報を示す概念図である。

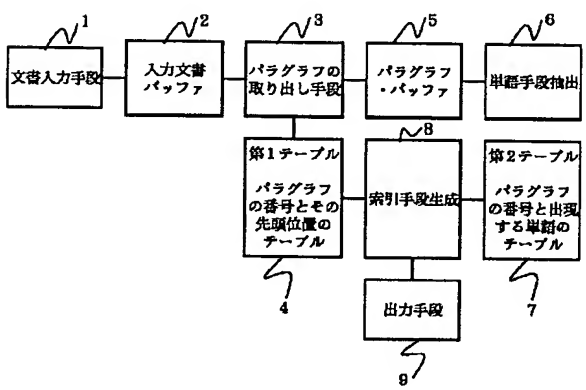
【図４】　本発明の検索用索引生成装置で生成される索引の一例を示す概念図である。

【図５】　従来の検索用索引の一例を示す概念図である。

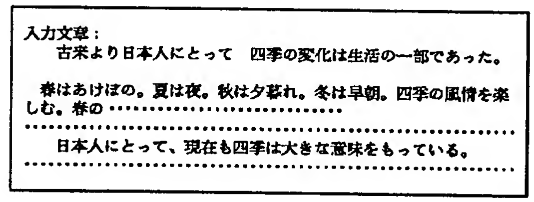
【符号の説明】

１・・・文書入力手段、　２・・・入力文書バッファ、　３・・・パラグラフ取出し手段、　４・・・第１テーブル、５・・・パラグラフバッファ、　６・・・単語抽出手段、７・・・第２テーブル、　８・・・索引生成手段、　９・・・出力手段、

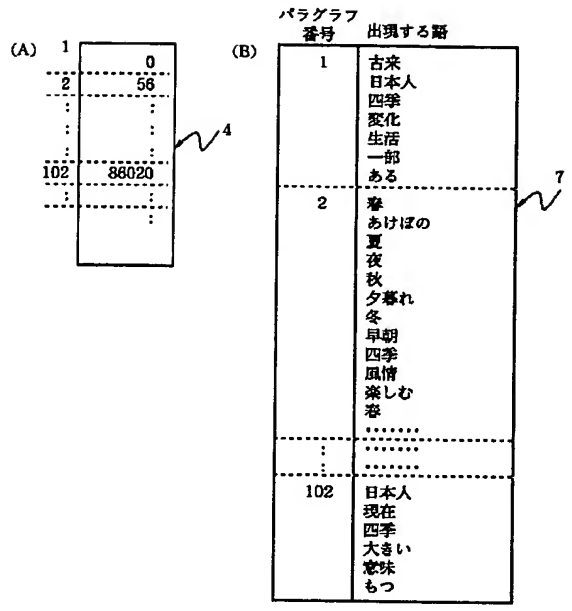
【図１】



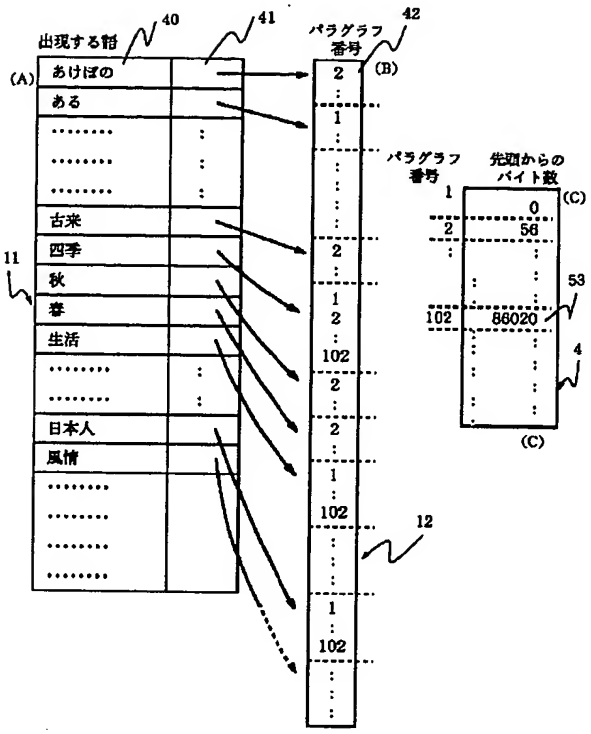
【図２】



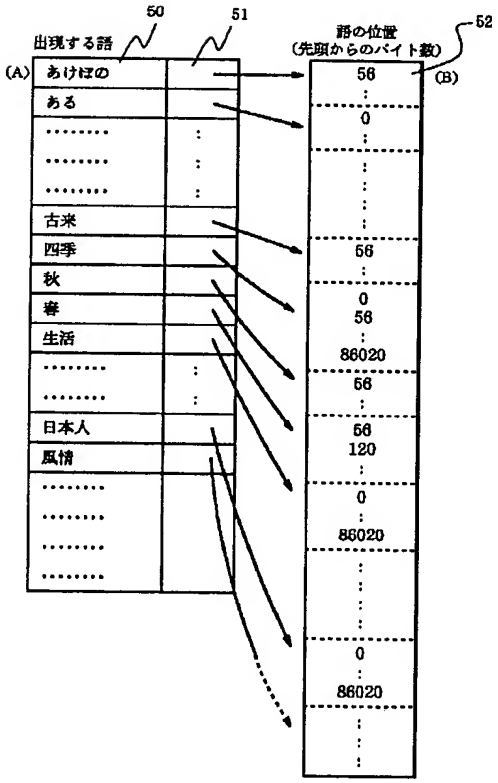
【図３】


(A)
1
0
2
56
102
86020
4
(B)
パラグラフ番号
出現する語
1
古来
日本人
四季
変化
生活
一部
ある
2
春
あけぼの
夏
夜
秋
夕暮れ
冬
早朝
四季
風情
楽しむ
春
102
日本人
現在
四季
大きい
意味
もつ
7

【図４】


出現する語
40
41
パラグラフ番号
42
(A)
(B)
あけぼの
ある
古来
四季
秋
春
生活
日本人
風情
11
12
パラグラフ番号
先頭からのバイト数
(C)
0
56
53
102
86020
4

【図５】


出現する語
50
51
語の位置（先頭からのバイト数）
52
(A)
(B)
あけぼの
ある
古来
四季
秋
春
生活
日本人
風情
56
0
56
86020
56
56
120
0
86020
0
86020